3D SHAPE FROM MULTI-CAMERA VIEWS BY ERROR PROJECTION MINIMIZATION

Gloria Haro

Universitat Pompeu Fabra (UPF)
Barcelona (Spain)

Montse Pardàs

Universitat Politècnica de Catalunya (UPC)
Barcelona (Spain)

ABSTRACT

Traditional shape from silhouette methods compute the 3D shape as the intersection of the back-projected silhouettes in the 3D space, the so called *visual hull*. However, silhouettes that have been obtained with background subtraction techniques often present miss-detection errors (produced by false negatives or occlusions) which produce incomplete 3D shapes. Our approach deals with miss-detections and noise in the silhouettes. We recover the voxel occupancy which describes the 3D shape by minimizing an energy based on an approximation of the error between the shape 2D projections and the silhouettes. The energy also includes regularization and takes into account the visibility of the voxels in each view in order to handle self-occlusions.

1. INTRODUCTION

Many applications in computer vision require the 3D reconstruction of a shape from its different views. When the available information in the images is just a binary mask segmenting the object the problem is called *shape from silhouette* (SfS). As first proposed by Baumgart [1] and then studied by Laurentini [2], the shape is usually computed as the maximum volume consistent with the given set of silhouettes, the so called *visual hull*.

Shape from silhouette techniques can be classified in three types: volume-based, surface-based or image-based approaches. Volume-based methods such as [3, 4] rely on a discretization of the 3D space. They suffer from quantization artifacts and require a lot of memory although they are robust and simple. Surface-based approaches [5, 6, 7] work with a mesh representation of the visual hull surface but usually develop numerical instabilities and often require a post-process of the obtained surface. Finally, inspired by the image-based visual hull [8], where geometric computations are done in the image space thanks to epipolar geometry, the work in [9] proposes to scan the 3D space with a geometry adapted to the images. The elementary unit of the proposed scanning geometry is the *conexel*, the 3D intersection of cones obtained by back projecting image quadrants.

GH performed this work while at UPC thanks to Juan de la Cierva program. Both authors thank project CENIT-VISION 2007-1007.

In real multi-view applications, silhouette masks are extracted from the different views by background subtraction techniques that segment foreground objects. These silhouettes often contain errors due to noise, calibration errors or background subtraction errors (false alarms and miss-detections due to similar foreground and background colors or even occlusions in the scene view). Some works have been proposed in order to deal with noisy silhouettes. For example, [10] is a volume-based approach which considers that a voxel projects to silhouette in a certain view if a high proportion of the pixels where the voxel is projected to are silhouette pixels. A method based on sensor fusion [4] considers each camera as an occupancy sensor and a probabilistic occupancy in the 3D space is inferred from the foreground probabilities in each view. It also handles situations where parts of the object of interest are outside the visible cone of some images, like [11].

However, very few attention has been payed to the case of silhouettes which present miss-detection errors. Traditional methods which reconstruct the shape just by computing the visual hull will produce incomplete shapes. If one wants to use the shapes for further applications like tracking or human body analysis by articulated models then these defective shapes are not desirable. A possible solution is to recover the shape as the 3D intersection of at least $N_c - e$ consistent cones (instead of N_c as in the visual hull), where N_c is the amount of cameras and e is the accepted amount of errors of a shape region in the different views. Its main drawback is that it produces enlarged shapes and risks to introduce some artifacts. Motivated by these problems, [12] proposes the *shape from inconsistent silhouette* (SfIS). In a first step it computes the visual hull, then it projects it to the different views and computes the errors, also called *inconsistent silhouettes*. Then, based on a minimization of a voxel miss-classification probability, it decides the minimum number of inconsistent silhouette cones intersections that have to be produced so that it can be determined that a voxel is part of the shape. This probability estimation is based on three priors (parameters): voxel occupancy probability, probability of silhouette false alarm and probability of silhouette miss-detection (assuming the two latter are constant for all views). The probability of a voxel miss-classification takes into account self-occlusions but only with respect to the voxels previously classified (by the visual hull step) as occupied. This lack of feedback and

the excessive number of parameters are not desirable.

A voxel-based method to recover a shape from incomplete silhouettes is proposed in this work. Inspired by [12], we define an energy based on the error between the silhouettes and the shape projections. The shape which best describes the set of silhouettes is the one that minimizes the energy. Voxel visibility in each view is also considered so that self-occlusions are allowed. Moreover, the energy is provided with a regularization term. The only parameter is the regularization factor. The energy is minimized with an iterative scheme based on an alternating optimization via gradient descent.

The paper is organized as follows. Section 2 explains our approach, based in an energy minimization. Section 3 refers to the main steps of the numerical optimization of the energy and some results are shown. Finally, conclusions and future work are presented in Section 4.

2. PROPOSED APPROACH

We propose a voxel-based approach to compute the 3D shape as the discrete occupancy function which minimizes an energy based on the error between the silhouettes and the 2D projections of the reconstructed shape. We also regularize the occupancy function via Total Variation in order to restrict the amount of solutions. Let Ω be a subset of $\mathbb{Z}^3$, the occupancy function $C : \Omega \rightarrow [0,1]$ indicates the probability, in each voxel $z \in \Omega$, of being shape (3D foreground). We denote by S_i the image containing the 2D silhouette of view i, where $i = 1, ..., N_c$, and N_c is the amount of cameras. Depending on the application, the function S_i will take values in $[0,1]$ (if it contains 2D foreground probabilities) or in $\{0,1\}$ (just a foreground mask). We consider the following energy:

$$E(C) = \alpha \sum_{z\in\Omega} |\nabla C(z)| + \sum_{z\in\Omega} \sum_{i=1}^{N_c} \frac{\mathcal{E}(S_i(P_i z), C)}{\mathcal{L}_i(P_i z)}, \quad (1)$$

where $\nabla C(z)$ is the discrete gradient at voxel z, $P_i z \in \mathbb{R}^2$ is the 2D projection of the center of voxel z in view i, and the function $\mathcal{E}(S_i(P_i z), C)$ is the error between the image value $S_i(P_i z)$ and the projected value $\hat{S}_i(P_i z, C)$ of shape C in $P_i z$: $\mathcal{E}(S_i(P_i z), C) = \left| S_i(P_i z) - \hat{S}_i(P_i z, C) \right|$. Notice that the error function is not summed over the image (data) domain, but over Ω instead, and so the same error will be counted several times. That is why the error expression is properly normalized by $\mathcal{L}_i(x)$, the amount of voxels in Ω which project onto x. Parameter α is related to the amount of noise in the silhouettes and will be fixed experimentally. We still need to give a precise definition for the projected silhouettes $\hat{S}_i(P_i z, C)$.

The projected image $\hat{S}_i$ contains 2D occupancy probabilities inferred from the 3D occupancy probabilities of voxels which fall in the line of sight of every point in the image plane. Since $C(z)$ is the occupancy probability of voxel z, then, the probability of being background is $1 - C(z)$. A point x in the projected image i will be background only if all the points in its line of sight, $l_i(x)$, are background. Then, assuming independence, the probability of x being background (bg) writes

$$P(x \in bg) = \prod_{k\in l_i(x)} (1 - C(k)).$$

Since $\hat{S}_i$ is the probability of projecting shape (foreground), $\hat{S}_i(x, C) = 1 - P(x \in bg)$ and the error expression is

$$\mathcal{E}(S_i(P_i z), C) = \left| S_i(P_i z) - 1 + \prod_{k\in l_i(P_i z)} (1 - C(k)) \right|. \quad (2)$$

However, the error expression (2) is non convex for $C \in [0,1]$. In the following, we are going to approximate the error (2) by a function whose second variation is positive and thus ensure the necessary conditions for a minimum. If we consider the particular case in which the image silhouettes are binary masks, that is $S_i \in \{0,1\}$, two cases can be distinguished in (2):

1. If $S_i(P_i z) = 1$, the error (2) reduces to

$$\mathcal{E}(S_i(P_i z), C) \mid_{S_i(P_i z)=1} = \prod_{k\in l_i(P_i z)} (1 - C(k)), \quad (3)$$

whose second variation is positive for $C \in [0,1]$.

2. If $S_i(P_i z) = 0$, the error reduces to a non convex function:

$$\mathcal{E}(S_i(P_i z), C)|_{S_i(P_i z)=0} = \left(1 - \prod_{k\in l_i(P_i z)} (1 - C(k)) \right). \quad (4)$$

The minimum of (4) is achieved when $C(k) = 0$ for all the $k \in l_i(P_i z)$. Then we can use the following equivalence:

$$\min \left(1 - \prod_{k\in l_i(P_i z)} (1 - C(k)) \right) \equiv \min \sum_{k\in l_i(P_i z)} C(k) \quad (5)$$

Using (5) and (3), the approximated error can be written as

$$\begin{aligned} \mathcal{E}(S_i(P_i z), C) = S_i(P_i z) & \prod_{k\in l_i(P_i z)} (1 - C(k)) \\ & + (1 - S_i(P_i z)) \sum_{k\in l_i(P_i z)} C(k). \end{aligned} \quad (6)$$

The error (6) consists of two complementary terms (errors), note that only one of the two is active at a time depending on the value of $S_i(P_i z)$, which acts as a 'switcher' between both. This error approximates $\mathcal{E}(S_i(P_i z), C)$ in (2) when the data images S_i are binary masks, but notice that the minimization

of (6) still makes sense in the more general case of probability images ($S_i \in [0,1]$). The new error function satisfies the necessary conditions for a minimum and the other term in (1) is convex. This error is not injective, but the regularization term in (1) will help in reducing the amount of possible solutions.

The energy just described assumes that a voxel z projects to a single point $P_i z$ in the image plane i. This is a coarse approximation for the discrete case. A voxel represents a rectangular prism which projects to a polygonal area having as vertices the eight projected voxel vertices [10]. The area and the shape of this polygon depends on the orientation and distance of the voxel with respect to the image plane. The size (area) of the polygon also depends on the voxel size (discrete resolution). The image is sampled in a rectangular lattice where the basic elements are the pixels and thus the voxel is not projected to a single pixel but to a collection of pixels that sample the projected polygon. We are going to see how we modify the energy E so as to refine the voxel projection to the image plane. In order to simplify computations, we will approximate a voxel by the sphere which circumscribes it so that the 2D projection is always a disc independently of the orientation of the image plane with respect to the voxel. Thus, a voxel projects to a set of pixels, sampling a disc, and the projected area (and as a consequence the amount of pixels in it) gets larger as the voxel size increases (lower 3D resolutions) and as the distance from the voxel to the image plane decreases. The energy (1) can be simply modified in order to include this new projection, we call the modified energy E_B,

$$E_B(C) = \alpha \sum_{z\in\Omega} |\nabla C(z)| + \sum_{z\in\Omega} \sum_{i=1}^{N_c} \sum_{x_j \in B_i(z)} \frac{\mathcal{E}(S_i(x_j), C)}{\mathcal{L}_i(x_j)\mathcal{B}_i(P_i z)}, \tag{7}$$

where $B_i(z)$ is the set of pixels falling inside the 2D ball centered at $P_i z$ with a radius inversely proportional to the distance from z to $P_i z$, and $\mathcal{B}_i(z)$ is the cardinality of this set.

Now consider the case of a voxel which is occluded, by other shape voxels, in some of the views. Those occluded views should not count in the error computation and as a consequence, the decision of shape/no shape in that voxel will be determined by a different amount of voting views, namely the visible views. We say that a view is occluded for a particular voxel if there are other voxels in the same line of sight which are set to '1' and are closer to the image plane. Following this definition, the visibility can be inferred from the occupancy values of the voxels in the same line of sight. In particular, the visibility of a voxel z in the view i at pixel x_j can be expressed as

$$V_i(z, x_j) = \prod_{\substack{k \in l_i(x_j) \\ d_i(k) < d_i(z)}} (1 - C(k)), \tag{8}$$

where $d_i(z)$ is the distance of voxel z to the i camera focal point. Then, the visibility depends only on the points in the same line of sight which are closer than z to the image plane. The visibility computed as (8) coincides with the probability that the points closer to the camera project as background. We propose to modify the energy (7) in order to take visibility into account. Then, the energy is also minimized with respect to the visibility functions V_i, $i = 1, ..., N_c$ and the visibility constraint (8) is included via a new term in the energy

$$E_B(C, V) = \sum_{z\in\Omega} \left(\alpha |\nabla C(z)| + \sum_{i=1}^{N_c} \sum_{x_j \in B_i(z)} \frac{V_i(z, x_j)\mathcal{E}(S_i(x_j), C)}{\mathcal{L}_i(x_j)\mathcal{B}_i(z)} \right) + \beta \sum_{z\in\Omega} \sum_{i=1}^{N_c} \sum_{x_j \in B_i(z)} \left(V_i(z, x_j) - \prod_{\substack{k \in l_i(x_j) \\ d_i(k) < d_i(z)}} (1 - C(k)) \right)^2,$$

where, by an abuse of notation, $\mathcal{L}_i(x_j)$ stands for the amount of visible voxels in the line of sight of pixel x_j. In this way, the projection error of a voxel in a view is *only* counted when the voxel under consideration is a visible voxel in that view. Minimizing this energy is equivalent to a maximum a posteriori estimation: The error term is the log-likelihood and the other two terms define the occupancy prior distributions.

3. RESULTS

The energy is minimized by an alternating optimization procedure where we first fix one of the variables of the energy and minimize with respect to the other by solving one step of the gradient descent and then we repeat the same process by interchanging the roles of variables. In practice, we consider a parameter β sufficiently large, and then we have a closed expression for the visibility (8). In order to reinforce the constraint $C \in [0,1]$ we clip the values that are not inside this interval.

We work with a synthetic shape from which we obtain 8 different views. The original silhouettes are shown in Figure 1(a). We have introduced some errors in the silhouettes of views 2 and 6, so as to evaluate how the proposed approach is able to recover the actual shape from incomplete silhouettes. The silhouette projections of the reconstructed shapes with the visual hull (VH) with N_c intersections, the VH with $N_c - 2$ intersections, the proposed method[1], and the shape from inconsistent silhouette (SfIS) are in figures 1(b) to 1(e) respectively. These projections have been computed by considering the sphere which circumscribes each voxel and projecting it to a disc with radius proportional to the distance to the camera. As it can be observed, the parts with e miss-detection in the silhouettes are recovered if we consider a VH with $N_c - e$ intersections but at the cost of enlarging the shape as $N_c - e$ decreases. The proposed approach works better than the SfIS and the VH reconstructions as it is able to recover the missing parts without enlarging the shape. Figure 2 shows the Hausdorff distance between the ground truth silhouettes and the re-

[1] A video with the 3D occupancy function is available at http://gps-tsc.upc.es/imatge/_Gloria/videos3D/wiamis09.htm.

(a) Original views

(b) Visual hull with N_c intersections

(c) Visual hull with $N_c - 2$ intersections

(d) Proposed approach ($\alpha = 0.5$)

(e) SfIS (probability parameters: $P_{BG} = 0.9$, $P_{MISS} = 0.1$, $P_{FA} = 0.1$)

Fig. 1. *Original silhouettes and projected silhouettes of reconstructions by different methods.*

covered silhouette projections by the different methods. Our approach gives the lowest distance in almost all the views.

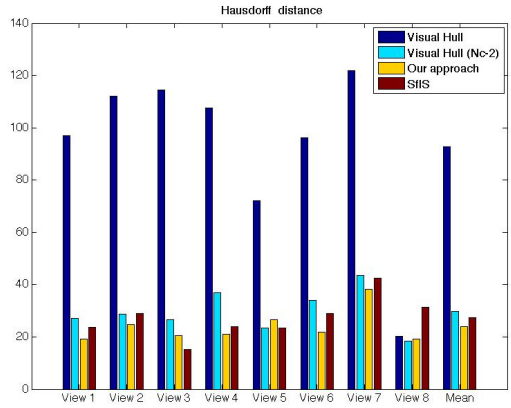

Fig. 2. *Comparison of Haussdorff distances to the ground truth silhouettes for each method. Each group of bars represents the results for the different views and the last group is the mean distance among the views.*

4. CONCLUSIONS

This paper studies the shape from silhouette problem when the silhouettes are incomplete in some of the views because of miss-detection errors or occlusions in the scene. For that, we have proposed the shape as the one which minimizes an energy based on the reprojection error plus a regularization term. The energy also takes the voxel visibility in each view into account in order to recover parts of the shape which are missed in some views and are occluded in others. Future work consists in optimizing computations and the evaluation in real scenarios.

5. REFERENCES

[1] B. G. Baumgart, *Geometric modeling for computer vision.*, Ph.D. thesis, Stanford University, 1974.

[2] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. on PAMI*, vol. 16, no. 2, pp. 150–162, 1994.

[3] D. Snow, P. Viola, and R. Zabih, "Exact voxel occupancy with graph cuts," in *Proc. of CVPR*, 2000.

[4] J. Franco and E. Boyer, "Fusion of multi-view silhouette cues using a space occupancy grid," in *Proc. of ICCV*, 2005.

[5] J. Franco and E. Boyer, "Exact polyhedral visual hulls," in *Proc. of the British Machine Vision Conference*, 2003.

[6] J. Franco and E. Boyer, "Efficient polyhedral modeling from silhouettes," *IEEE Trans. on PAMI*, vol. In press, 2008.

[7] W. Matusik, C. Buehler, and L. Mcmillan, "Polyhedral visual hulls for real-time rendering," in *Proc. of Twelfth Eurographics Workshop on Rendering*, 2001, pp. 115–125.

[8] W. Matusik, C. Buehler, R. Raskar, L. McMillan, and S. Gortler, "Image-based visual hulls," in *Proc. of SIGGRAPH*, 2000.

[9] J. R. Casas and J. Salvador, "Image-based multi-view scene analysis using 'conexels'," in *Proc. of the HCSNet workshop on Use of vision in human-computer interaction*, 2006.

[10] K. M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler, "A real time system for robust 3d voxel reconstruction of human motions," in *Proc. of CVPR*, 2000.

[11] B. Michoud, E. Guillou, and S. Bouakaz, "Shape from Silhouette: Towards a solution for partial visibility problem," in *Proc. of Eurographics*, 2006, pp. 13–16.

[12] J. L. Landabaso, M. Pardàs, and J. R. Casas, "Shape from inconsistent silhouette," *CVIU*, vol. In press, 2008.